Report on

Visualization and exploration of Adult Dataset

CS245, Data Science Foundation, Chaojun Lu, Autumn 2017

叶泽林 515030910468

1 Introduction

Currently, data science is becoming ubiquitous in our society and showing its essentiality in many domains (e.g. medical industry [1, 2], finance [3], social media [4, 5]). With the rapid evolution and wide applications of data science, a series of efficient packages have been constantly developed [6, 7, 8, 9]. Utilizing these packages skillfully is of great importance nowadays. In this project, I tend to conduct an exploration over the *Adult* dataset and visualize the results with some of these packages.

2 Approaches

The *Adult* dataset is in the format of .CSV with 16281 lines, each line contains some basic information (age, job, gender and etc) of an adults. I explore and extract the information hidden in the data with the following steps:

1. Read the data with *pandas* [7] and reconstruct it as *DataFrame* type.
2. Select a target item (e.g. work time per week) to conduct analysis.
3. Select two items to explore the relation between them.
4. Visualize all analysis results with *matplotlib* [8] and *pandas*.

3 Experiments

3.1 The Distribution of Single Item

3.2 The Relation between Two Items

4 Conclusion and Discussion

In this project, I carry out the visualization and exploration of *Adult* dataset and discover some interesting phenomenons reflected from it.

According to the results of above analysis, the distribution of each item has some kinds of ralation with others. Thus, I think I can train a model to predict target items from other items in my future learning process.

Ultimately, I tend to express my sincere thanks to Professor Chaojun Lu for his patient explanation and guidance in lectures! Thank you!

References

[1] S. P. Bhavnani, D. Muñoz, and A. Bagai, “Data science in healthcare: implications for early career investigators,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 683–687, 2016.

[2] Y. Liang and A. Kelemen, “Big data science and its applications in health and medical research: Challenges and opportunities,” *Austin Journal of Biometrics & Biostatistics*, vol. 7, no. 3, 2016.

[3] S. O’ Halloran, S. Maskey, G. McAllister, D. K. Park, and K. Chen, “Data science and political economy: application to financial regulatory structure,” *RSF*, 2016.

[4] W. Xiao-fan, “Data science and social network: Big data, small world,” *Science and Society*, vol. 1, p. 003, 2014.

[5] C. Vande Kerckhove, *Data science for modeling opinion dynamics on social media*. PhD thesis, UCL-Université Catholique de Louvain, 2017.

[6] “Numpy.” http://www.numpy.org/.

[7] “Pandas.” http://pandas.pydata.org/.

[8] “matplotlib.” https://matplotlib.org/.

[9] “scikit-learn.” http://scikit-learn.org/stable/.